Mask-Ranking Network for Semi-Supervised Video Object Segmentation

Wenjing Li, Xiang Zhang(✉), Yujie Hu, and Yingqi Tang

University of Electronic Science and Technology of China, Chengdu, China
{liwenjing,huyujie,tangyingqi}@std.uestc.edu.cn
uestchero@uestc.edu.cn

Abstract. Video object segmentation is the fundamental problem of video analysis and many methods based on mask propagation and matching have been proposed in recent years. However, the two strategies are highly dependent on the last mask or the fixed mask given in the first frame and hence cannot adapt well to high deformation and rapid motion of objects. In this paper, we proposed a novel architecture named Mask-Ranking Network(MRNet), which takes advantage of both the propagation-based method and the matching-based method, to address the above problem. Specifically, in order to make better use of the long-term previous masks, we propose a novel propagation mechanism to make the network comprehensively consider the previous information. Under a unified encoder-decoder framework, we track the pixel-wise similarity of the object activation area in a long-term manner and explore the correlation between frames. In contrast to propagation-based only or matching-based only techniques, our method reduces the accumulation of errors in the propagation process and effectively uses the long-term previous frame information. In the video object segmentation task, MRNet can better handle the deformation of the objects, and make the segmentation result more accurate. We validate the effectiveness of the proposed method on the DAVIS 2016 and DAVIS 2017 dataset. Experiment results show that our method achieve state-of-the-art performance without using online fine-tuning and is robust to long-term propagation.

1 Introduction

Video Object Segmentation(VOS) aims to separate the object(s) of interest from the background pixels throughout a given video. With the rapid development of deep learning in recent years, as the basis of video analysis and subsequent video processing, this fundamental task has been applied to various fields, such as scene analysis, autonomous driving, action recognition and so on. In the aspect of setup, two main types of this problem are unsupervised and semi-supervised which differ from each other in whether the object annotation(s) in the first frame of the video is provided. In this paper, we consider the semi-supervised setting, in which the groundtruth segmentation of one or multiple objects are given in the first frame of the video, and then the methods automatically estimate the segmentation results in the rest of video. However, even with some prior

knowledge in the process of inference, this is still a challenging task because the appearance of the object can drastically change throughout the video due to the deformation, occlusion and illumination change, greatly deteriorated the segmentation results.

To tackle the aforementioned challenges, many algorithms based on deep learning have been proposed in recent years. Some algorithms regard the VOS as a mask-refinement process, which belongs to the mask-propagation based method. This type of algorithm uses a neural network to learn the deformation from the previous output to the query frame, starting from the first frame. These networks learn the features of the previous mask prediction and adjust it to fit the current frame, usually are simple in structure and performing well on the smooth deformation. However, they are susceptible to rapid motion and suffer from error accumulation during propagation. Another type of method focuses on finding the connection between the query frame and the first frame, which belongs to the matching-based method. The standard strategy is to extract the features of the current frame and the reference through neural network and do high-dimensional pixel-wise metric matching. These methods avoid the loss of information in the mask propagation process, but directly using the k-nearest neighbour results as the final classification makes the segmentation rougher.

In this paper, we proposed a novel neural network Mask-Ranking Network (MRNet) for semi-supervised VOS task that integrate the advantages of both propagation-based and matching-based methods. We conduct a Mask-Ranking Module (MRM), to dynamically and rapidly select the most conductive mask to guiding the segmentation in the intermediate process of mask-propagation. Through MRM, the network can not only avoid the absoluteness of the hard classification based on the matching method, but also continuously modify the propagation results during the inference process. With our framework, the network is no longer limited to relying only on the first frame annotation and the previous output, as the conductive information can be easily added. The proposed network is also highly efficient as there is no need of fine-tuning in the test time, which is a truly end-to-end training network.

The major contributions of this paper are: (i)We proposed a novel network Mask-Ranking Network (MRNet) for semi-supervised video object segmentation, which can easily and continuously add conductive information to refine the segmentation result. (ii)We conduct a Mask-Ranking Module to dynamically and rapidly guiding the segmentation in the intermediate process of inference. (iii)Experiments on DAVIS 2016 and DAVIS 2017 show that the proposed method exceed the state-of-the-art performance.

2 Related Work

2.1 Unsupervised and Semi-Supervised Video Object Segmentation

Video object segmentation can be divided into two types: unsupervised and semi-supervised, which differ from each other in whether the first frame groundtruth

is provided. Methods based on unsupervised setting [1,2,3,4] mainly explore the dense optical flow and object appearance features to do pixel-wise prediction. However, the object of interest is not specified in the video, making the segmentation result is ambiguous. In this paper, we focus on the semi-supervised algorithms. Many semi-supervised approach rely on fine-tuning on the first frame during testing in order to obtain better performance [5,6,7,8,9,10,11,12,13]. It has shown that fine-tuning on the first frame significantly improves the accuracy. OSVOS [6] fine-tunes the pre-trained convolutional network on the annotated first-frame at test time. OnAVOS [12] and OSVOS-S [11] are build on the OSVOS. OnAVOS employs an online adaption mechanism by treat the segmentation results as new training examples to update the network online during the test time. OSVOS-S transfer the generic semantic information learned on ImageNet to the segmentation task based on a fully convolutional neural network. However, the expensive computation and time-delay extremely limit the real-time processing applications. Therefore, there are several recent works aim to achieve a better run time and usability by avoiding online learning [14,15,16,17,18]. FRTM [17] is composed of appearance model and segmentation model. The target adaption process of FRTM is fully simulated during the offline training stage. SAT [18] treats each target object as a tracklet, avoiding the effect of online fine-tuning to achieve real-time segmentation. In this paper, we explore an end-to-end network structure, which dynamically selects necessary information during the intermediate process of forward propagation to improve the segmentation effect, completely avoiding online training.

2.2 Matching-based Methods

Matching-based method is to exploit the appearance similarity between the current frame and reference frame. They directly segment each pixel based on the result of pixel-level matching [19,20,21,22,23,24,25,26,27,28,29]. PML [25] treats the video object segmentation as pixel-wise retrieval problem. They first trained an embedding layer with triplet loss and then predict each pixel by nearest-neighbour matching result to the first frame. However, this type of hard classification often results in noisy segmentation. VideoMatch [26] adopts a soft matching mechanism which is similar to PML. It uses a soft matching layer to produce the foreground and background similarity maps and consider the k nearest neighbours of each pixel for segmentation. However, the result is still derived from the matching score, making the segmentation unsatisfactory. FEELVOS [28] uses an embedding layer to calculate global and local pixel-wise matching in the internal of network, but suffering the lose of the information from similarity maps due to the propagation. RANet [27] learns pixel-wise similarity maps to explore the similarity between the first frame and the current frame by a ranking attention module. Recently, Zhang et al. [29] proposed a transductive method TVOS, that takes a label propagation methods where the labels are propagate based on feature similarity in an embedding space. Different from the previous works, we tried to make use of the information of more previous frames, instead of just

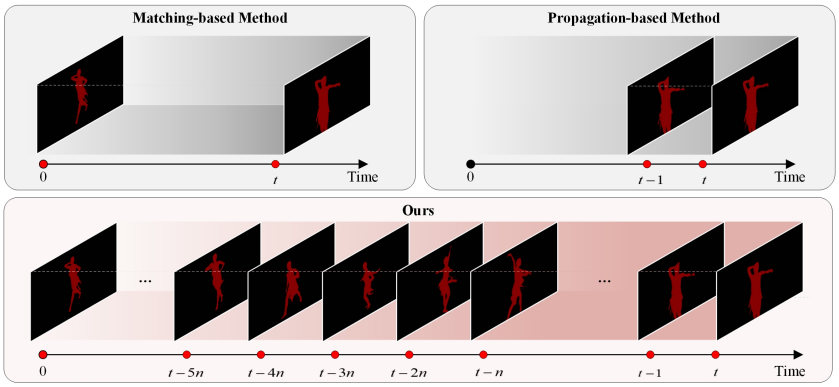

Fig. 1: A simple comparison of propagation-based methods, matching-based methods and our framework. The time marked with red dots represents the information at that moment is used to guide the segmentation of current frame.

selecting the first frame as guidance or directly taking the nearest neighbour matching as the final segmentation decision.

2.3 Propagation-based Methods

The propagation-based methods [30,31,29,32,33,3,34,35,36,37,38,39] mainly rely on the segmentation of the previous frame to improve the performance of the current frame. VPN [40] propagates structure information through the entire video by a unified framework of temporal bilateral network and spatial refinement network. OSMN [41] combines a segmentation network with two modulators, which manipulate the intermediate layers of the segmentation network and learning the annotated first frame and spatial location of the previous frame, respectively. RGMP [34] constructs the network as a Siamese encoder-decoder structure, in which the weights of the encoder part are shared between two streams. One of the encoder stream takes the current frame with previous mask estimation as input. The other stream takes the first frame and its annotation as input. The architecture of RGMP is similar with ours, as we also utilize the Siamese network in the encoder part. However, instead of simply stacking the feature maps, we dynamically select the most conductive mask by proposed Mask-Ranking Module in the intermediate process and feed it into the decoder. Recently, TVOS [29] uses the previous segmentation as training data for discriminative model. The mask of the previous frame is feed into the target module to generate the low-resolution score map during the inference process. RANet [27] uses a ranking attention module to filter the similarity maps and then feed them together with the previous mask estimation into the decoder, which makes it easier for the network to capture useful information. However, directly feeding the previous segmentation into the decoder can easily lead to the error accumulation.

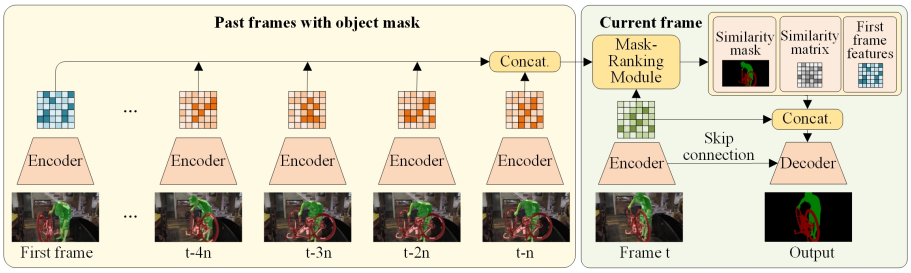

Fig. 2: An overview of our architecture. The network contains two encoders that encode the past frames and the current frame, respectively. The features of past frames(in orange) are concatenated, and then passed into the matching layer to calculate the correlation with the feature of the current frame(in green). Finally, the similarity mask, similarity matrix(in grey) and the first frame features(in blue) is concatenated and passed into decoder together with current frame features that skip connection.

3 Method

Given the annotation of the first frame, many previous semi-supervised video object segmentation methods mainly explore the relationship between the current frame and the previous or first frame. In this paper, we proposed a novel architecture named Mask-Ranking Network(MRNet) for semi-supervised video object segmentation task. The motivation of our method is illustrated in Fig. 1. The key idea of our method is to explore more information from all of the previous frames while without any online learning. Our method can constantly update the content of reference to make it more consistent with the current frame. It combines the advantages of the propagation-based method and the matching-based method, so that the network can constantly correct errors during the inference process and be robust to long-term videos.

In this section, we first provide an overview of the proposed MRNet in § 3.1. In § 3.2, we describe the proposed Mask-Ranking Module in details. In § 3.3, we discuss the details during the inference. Finally, the extension for multi-object is presented in § 3.4.

3.1 Overview of the architecture

Our MRNet can be divided into two parts: the part processing the past frames and the part processing the current frame. An illustration of our MRNet is shown in Fig. 2.

Processing the past frames. In the left part of Fig. 2, we exhibit the processing of a series of past frames. Each encoder takes a RGB image and mask as input and the weights of the encoders are shared. Among them, the first frame and its annotation are the reference frame and reference mask of this video, respectively. For other past frames, they are combined with the mask of their

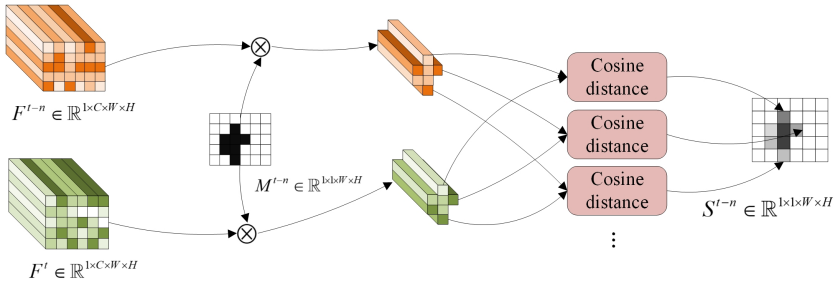

Fig. 3: An illustration of the process of mask similarity matching. $F^t \in \mathbb{R}^{1\times C\times W\times H}$ represents the features of the current frame at time t. $F^{t-n} \in \mathbb{R}^{1\times C\times W\times H}$ represents the features of the past frame at time $t-n$. $M^{t-n} \in \mathbb{R}^{1\times 1\times W\times H}$ represents the mask of the frame at time $t-n$. $S^{t-n} \in \mathbb{R}^{1\times 1\times W\times H}$ represents the similarity matrix between the current frame and the past frame at time $t-n$.

previous frame as the input of the encoder. In order to reduce the computation cost, the past frames are selected at an interval n. After getting the features of all the past frames, we concatenate all of the features and passed them to the matching layer for subsequent matching and selection. Specifically, we choose ResNet50 [42] as the shared feature extractor. The number of input channels has been adjusted to 4 to receive the input we set, containing 3 channel RGB image and 1 channel mask. The weights of the network are initialized from the ImageNet pretrained model.

Processing the current frame. The current video frame is processed by an encoder-decoder structure on the whole. The inputs of encoder are an RGB image and the mask from the previous frame. Features extracted by the backbone are stored and fed into the Mask-Ranking Module to learn the similarity with the past frame features. The output of Mask-Ranking Module contains three parts, the similarity mask, similarity matrix and the stored first frame features, which we will discuss in details in the 3.2. Then, the three outputs are concatenated with the current frame high-dimensional features and passed through our pipeline. The decoder consists of three refinement modules [43], which take the concatenated features and the skip connections features as input. What is different from the original architecture is that the convolution layers have been changed into residual blocks [44]. We first up-sample the deep layer features of by bilinear interpolation, and then concatenate them with shallow features of current frame from the encoder. In this way, the features of different depths in the encoder are fed into the decoder to obtain a more comprehensive segmentation result.

3.2 Mask-Ranking Module

The propagation-based method uses the previous frame to guide the segmentation of the current frame, and the matching-based method explores the relation-

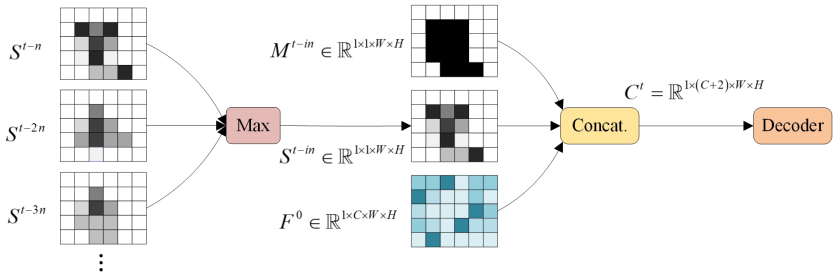

Fig. 4: An illustration of the process of mask ranking. After computing the similarity matrix, we choose the mask with the highest similarity matrix value. We choose the mask with the largest similarity matrix value, and feed it to the decoder with the similarity matrix and the mask of the first frame. S, M and the F^0 denote the similarity map, mask and the first frame feature, respectively. C^t represent s the concatenate features, with a dimension of $1 \times (C+2) \times W \times H$.

ship between the current frame and the first frame. However, just propagating the mask throughout the video without any refinement will lead to a poor performance in the long-term video. And the deformation of the object makes it difficult for the network to perform an accurate pixel-wise binary classification just through the similarity with the first frame. In order to use the features from the semantically related region in the past frames to help refine the current frame segmentation, we propose a Mask-Ranking Module, which synthesizes the advantages of the propagation-based and matching-based video segmentation methods. There are mainly two operations in our Mask-Ranking Module, one is the process of mask similarity matching based on the object activate region, and the other is the mask ranking operation to rank and select the previous masks and similarity matrices.

Mask similarity matching. A schematic illustrating details of the process of mask similarity matching is given in Fig. 3. Different from other matching-based methods that based on Euclidean distance to obtain pixel-wise similarity, we choose cosine similarity for our mask similarity matching. For a given object, we match the semantic activate region features between the current frame and the past frames to obtain the similarity matrices. Denote the current frame is at time t, the past frame features is defined as:

$$F = \left\{F^0, F^1, \cdots, F^{t-1}\right\} \tag{1}$$

Due to the similarity between adjacent frames, we balance the amount of calculation and the amount of information in past frames, and choose to perform mask similarity matching every n frames. In order to ensure a fixed number of channels, k feature maps are selected for calculation each time. Then, the features to be matched are $\left\{F^{t-kn}, \cdots, F^{t-2n}, F^{t-n}\right\}$. The masks of the selected features are $\left\{M^{t-kn}, \cdots, M^{t-2n}, M^{t-n}\right\}$. We denote the foreground feature vector at location (i, j) of the F^{t-kn} as m^{ij}_{t-kn}. The feature vector at the corresponding

position (i, j) of the current frame feature map F^t is x_t^{ij}. The similarity matrix between the current frame and the past frame at time $t - kn$ is denoted as $S^{t-kn} \in \mathbb{R}^{1\times1\times W\times H}$. We denote the foreground similarity value at location (i, j) as s_{t-kn}^{ij}. The background pixels similarity value is set to 0. The cosine similarity between the two feature vectors m_{t-kn}^{ij} and x_t^{ij} is formulated as:

$$s_{t-kn}^{ij} = \frac{m_{t-kn}^{ij} x_t^{ij}}{\left\| m_{t-kn}^{ij} \right\| \left\| x_t^{ij} \right\|} \tag{2}$$

Mask ranking. The process of mask ranking is shown in Fig. 4. After the calculating of the similarity matrix between the past frame features $\{F^{t-kn}, \cdots, F^{t-2n}, F^{t-n}\}$ and the current frame feature F^t, we rank the similarity matrix according to the value sum of them. The similarity matrix is obtained by cosine similarity between two vectors m_{t-kn}^{ij} and x_t^{ij}, where a larger value represents the higher the similarity of the feature vectors at the pixel. The similarity matrices are denoted as $\left\{S^{t-kn}, \cdots, S^{t-2n}, S^{t-n}\right\}$, and the dimension is $1 \times 1 \times W \times H$. The value sum of S^{t-kn} is calculated as:

$$V_{t-kn} = \sum_{i=1}^{H} \sum_{j=1}^{W} s_{t-kn}^{ij} \tag{3}$$

However, if we just make decision by the value V_{t-kn}, a similarity matrix with a large number of foreground pixels is easier to be selected because its similarity matrix value may be relatively larger. So, we use the average of V_{t-kn} to rank:

$$averageV_{t-kn} = \frac{1}{\sum_{i=1}^{H} \sum_{j=1}^{W} m_{t-kn}^{ij}} V_{t-kn} \tag{4}$$

We denoted the time with the largest $averageV_{t-kn}$ as $t - in$. The similarity matrix S^{t-in}, similarity mask M^{t-in} and the first frame feature map F^0 are concatenated to the feature $C^t \in \mathbb{R}^{1\times(C+2)\times W\times H}$ and fed into the decoder.

3.3 Inference

The inference of the proposed MRNet is straight forward in an end-to-end manner. Given the first frame annotation, the network can automatically propagate the mask throughout the video with the dynamic mask ranking internal operation. When we segment the frame at time t, k feature maps are selected from the past frames according to the time interval n to learn the similarity with the current frame. Then the similarity matrix and mask with the largest similarity matrix $averageV_{t-kn}$ value are selected to refine the segmentation result of the current frame. Every feature map is calculate only once in the pipeline, making the technique efficient enough. In particular, two frames with long time intervals have low similarity, and two frames with short time intervals are usually similar.

Table 1: Past frames selection criteria

Current frame time	Selected past frames time
$t \leq 5$	0
$5 < t \leq 10$	0, $t-5$
$10 < t \leq 15$	0, $t-5$, $t-10$
$15 < t \leq 20$	0, $t-5$, $t-10$, $t-15$
$t > 20$	0, $t-5$, $t-10$, $t-15$, $t-20$

In order to avoid redundancy of information and waste of calculation, the k is set to 4 and the n is set to 5 in our implementation. In this way, if the time t of the current frame is greater than 20, the interval reference time for past frames is 20 frames. When t is less than 20, we increase the available past frames as the video passed through the network frame-by-frame. Specifically, if $t \leq 5$, we only consider the first frame; if $5 < t \leq 10$, we consider the first frame and the frame $t-5$; if $10 < t \leq 15$, we consider the frame $t-5$, $t-10$ and the first frame; if $15 < t \leq 20$, we consider the frame $t-5$, $t-10$, $t-15$ and the first frame. The past frames selection criteria is shown in Table. 1.

3.4 Extension for Multi-object VOS

The extension from single-object video object segmentation to multi-object video object segmentation is to run each object independently. In the proposed MRNet, we individually match the feature vector of each object between the past frames and current frame. As for N objects in a frame, in order to keep the the number of output channels of the Mask-Ranking Module unchanged, we unify the N similarity matrix of each object into one. Specifically, taking two objects as an example, the final similarity matrix $A \in {}^{1\times1\times W\times H}$ is the average of the two similarity matrices $S^{ij}_{t-kn,1}$ and $S^{ij}_{t-kn,2}$. Each pixel value is calculated as:

$$A^{ij} = \frac{1}{2}\left(S^{ij}_{t-kn,1} + S^{ij}_{t-kn,2}\right) \tag{5}$$

The subscript $t-kn$ of $S^{ij}_{t-kn,N}$ represents the frame index, and the subscript N represents the object index in the frame.

4 Experiments

4.1 Implementation Details

Training datasets. The DAVIS 2016 [45] dataset is for single-object segmentation which contains a total of 50 sequences, 3455 frames with densely pixel-wise annotations. The 50 sequences are divided into a training set with 30 sequences, and a validation set with 20 sequences. The DAVIS 2017 [46] dataset is the extension of multi-object segmentation, which contains a training set with 60

sequences and a validation set of 30 videos. Each sequence from DAVIS 2016 and DAVIS 2017 has a temporal extent about 2-4 seconds, that all major challenges in longer video sequences are included.

Network settings and training details. We use the ResNet50 [42] as the encoder feature extractor, and the parameters of it is initialized by a pre-trained model on ImageNet. The weights of the entire network are shared between process of the past frames and the current frames. The channels of the encoder output are 2048, and the input channels of the decoder are $(2048 \times 2 + 2)$, which contain the channels of the current frame features, the first frame reference features, selected similarity matrix and mask. During the training process, we adopt the thought of BPTT [47], the length of it is taken as 12. The entire network runs on a single NVIDIA GeForce GTX 1080Ti GPU and is trained end-to-end using the Adam optimizer [48]. The weight decay factor is 0.0005 and the initial learning rate is set to 10^{-5} and gradually decreases overtime.

Evaluation metrics. Following the suggestion of DAVIS [45], we use three standard metrics: the region similarity $\mathcal{J}$ Mean, the contour accuracy $\mathcal{F}$ Mean and $\mathcal{J}\&\mathcal{F}$, which is the average of $\mathcal{J}$ Mean and the $\mathcal{F}$ Mean. The Jaccard index $\mathcal{J}$ mean is calculated as the mean of the intersection over union($mIoU$) between the network output M and the groundtruth G, thus the metric represents the region similarity. As for the metric $\mathcal{F}$ mean, we consider it as a good trade-off between the $Precision$ and the TP_{rate}. The definitions are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$TP_{rate} = \frac{TP}{TP + FN} \tag{7}$$

$$\mathcal{F} = \frac{2 \times Precision \times TP_{rate}}{Precision + TP_{rate}} \tag{8}$$

$$\mathcal{J} = \frac{M \cap G}{M \cup G} \tag{9}$$

where TP, FP and FN are the numbers of true positives, false positives and false negatives, respectively.

4.2 Ablation study

We perform an extensive ablations on DAVIS 2016 validation set to confirm the effectiveness of our proposed method. In our ablative experiments, we first analyze the impact of proposed Mask Ranking Module(MRM) by totally removing it from our network. Then we verify the effectiveness of Similarity Matrix by leave its calculation out of decoder. Finally, we conduct a experiments on the impact of k and n. Ablation study results are shown in Table 6.

The Effectiveness of the Mask Ranking Module. Firstly, we remove the proposed Mask Ranking Module to analyze its impact, which leads to a dramatic reduction that the $\mathcal{J}\&\mathcal{F}$ reduce from the 85% to 81.1%, which is shown in the

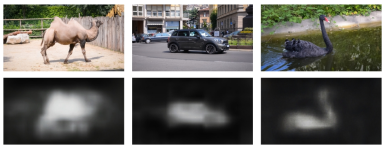

Fig. 5: Visualization of Similarity matrix on DAVIS 2016 dataset.

	$\mathcal{J}\&\mathcal{F}$
Baseline	81.1
+MRM(without Sim-map)	82.6
+MRM(with Sim-map)	85.0

Fig. 6: Ablation study of the proposed network on DAVIS 2016 dataset.

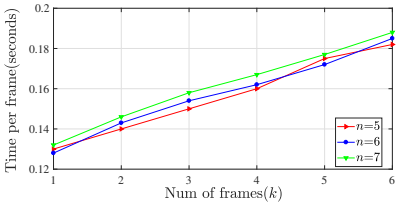

Fig. 7: The impact of k.

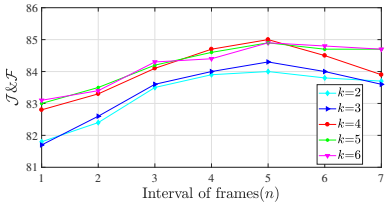

Fig. 8: The impact of n.

first row and the bottom row in Table 6. This results clearly demonstrate that the Mask Ranking Module we proposed plays an important role in our framework.

The Effectiveness of the Similarity Map. As for the effectiveness of the similarity matrix, we leave its calculation from the decoder, that the inputs of decoder only include the current frame feature, the first frame feature, the similarity mask and the skip-connection features from the encoder. As shown in the second row in Table 6, the $\mathcal{J}\&\mathcal{F}$ reduce from the 85% to 82.6%, and only 1.5% higher than baseline. This results demonstrate that the similarity matrix made a major contribution to the total accuracy. The visualization of the Similarity Matrix is shown in Fig 5, we can find that the Similarity Matrix can provide conductive information for segmentation.

The Impact of the k and n. The larger the n, the longer the time we consider, but the object may have a larger deformation during this time, which is of little guiding significance for current frame. The larger the k, the larger number of similar frame to be calculated, and the greater the amount of calculation. Therefore, we conducted experiments for choosing k and n. Fig 7 shows the impact of k. We can see that the time process one frame go longer as the k increases. Fig 8 shows the impact of n, and we can see that when n increase, the $\mathcal{J}\&\mathcal{F}$ first increases and then decrease. That is because the similarity become smaller when two frames are too far apart.

4.3 Comparison to the state-of-the-art

Comparison methods. We compare our MRNet with a total of 20 methods that contains 9 online based methods(OSVOS [6], MaskRCNN [33], Lucid [5], OSVOS-S [11], CINM [32], SegFlow [7], MSK [10], OnAVOS [12], PReMVOS [35]) and 11 offline methods(Videomatch [26], PML [25], VPN [40],

Table 2: The quantitative comparison on the DAVIS 2016 and DAVIS 2017 validation sets. The results are sorted for online(OL) and non-online methods respectively. The highest scores in each category are highlighted in bold.

Method	OL	DAVIS 2016				DAVIS 2017		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$ Mean	$\mathcal{F}$ Mean	Time	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$ Mean	$\mathcal{F}$ Mean
OSVOS [6]	✓	80.2	79.8	80.6	9s	60.3	56.6	63.9
MaskRCNN [33]	✓	80.8	80.7	80.9	-	-	60.5	-
Lucid [5]	✓	83	83.9	82.0	>100s	66.6	63.4	69.9
OSVOS-S [11]	✓	86.6	85.6	87.5	**4.5s**	68.0	64.7	71.3
CINM [32]	✓	84.2	83.4	85.0	>30s	70.7	67.2	74.2
SegFlow [7]	✓	75.4	74.8	76.0	7.9s	-	-	-
MSK [10]	✓	77.6	79.7	75.4	12s	-	-	-
OnAVOS [12]	✓	85.5	86.1	84.9	13s	67.9	64.5	71.2
PReMVOS [35]	✓	**86.8**	**84.9**	**88.6**	32.8s	**77.8**	**73.9**	**81.7**
Videomatch [26]	✗	80.9	81.0	80.8	0.32s	62.4	56.5	68.2
PML [25]	✗	77.4	75.5	79.3	0.28s	-	-	-
VPN [40]	✗	67.9	70.2	65.5	0.63s	-	-	-
OSMN [41]	✗	73.5	74.0	72.9	0.13s	54.8	52.5	57.1
FEELVOS [28]	✗	81.7	80.3	83.1	0.5s	69.1	65.9	72.3
RGMP[34]	✗	81.8	81.5	82.0	0.13s	66.7	64.8	68.6
A-GAME [30]	✗	82.1	82.0	82.2	0.07s	70.0	67.2	72.7
FAVOS [15]	✗	81.0	82.4	79.5	1.8s	58.2	54.6	61.8
TVOS [29]	✗	-	-	-	-	72.3	69.9	74.7
SAT [18]	✗	83.1	82.6	83.6	**0.03s**	72.3	68.6	76.0
FRTM [17]	✗	81.6	-	-	0.07s	69.2	-	-
MRNet(ours)	✗	**85.0**	**85.1**	**84.9**	0.16s	**73.4**	**70.4**	**76.3**

OSMN [41], FEELVOS [28], RGMP [34], A-GAME [30], FAVOS [15], TVOS [29], SAT [18], FRTM [17]).

Results on DAVIS 2016. Table 2 compares our methods on the DAVIS 2016 validation set to other state-of-the-art methods. Our MRNet achieves a $\mathcal{J}\&\mathcal{F}$ Mean of 85%. Among all the methods without OL techniques, the performance of the proposed MRNet is the best. Considering all the methods listed in Table 2, the online learning based method PReMVOS [35] has a $\mathcal{J}\&\mathcal{F}$ Mean of 86.8%, which is 1.8% higher than our MRNet. However, the time processing one frame of 32.8s is much longer than MRNet of 0.16s. Using additional traning data and employing online learning lead to a low processing speed of PReMVOS. While our MRNet avoid the online learning operation and post-processing, obtaining a efficient performance on the DAVIS 2016 validation set.

Results on DAVIS 2017. For the task of multi-object segmentation, we evaluate our method on DAVIS 2017 validation set. Table 2 shows a comparison to other state-of-the-art methods. Experiments shows that our MRNet achieves a $\mathcal{J}\&\mathcal{F}$ Mean of 73.4%, which is the best among all of the methods without

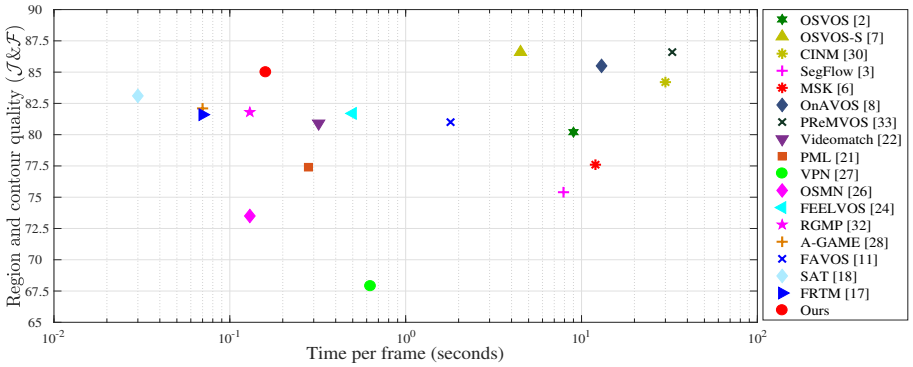

Fig. 9: A comparison of performance and speed for semi-supervised video object segmentation on the DAVIS 2016 validation set. The better methods are located at the uppper-left corner. The proposed MRNet shows a good speed/accuracy trade-off.

OL technique, demonstrating the superiority of the proposed MRNet on the multi-object segmentation.

Speed. A comparison of performance and speed for semi-supervised video object segmentation on the DAVIS 2016 validation set is shown in Fig. 9. The horizontal axis represents the time it takes the network to process a frame, and the vertical axis represents the metric $\mathcal{J}\&\mathcal{F}$ Mean. The method with less processing time and higher accuracy is superior. Therefore, the better methods are located at the uppper-left corner. The proposed MRNet shows a good speed/accuracy trade-off.

4.4 Qualitative result

Fig. 10 shows qualitative results of our MRNet on DAVIS 2016 and DAVIS 2017 validation set. It can be seen in many cases such as appearance changes, object fast motion, occlusion and so on, MRNet is able to produce accurate and robust segmentation results. While in some cases, the segmentation result of MRNet is not complete, this may be because the similar masks found by cosine similarity are not the most similar to the current frame in the global.

5 Conclusion

In this paper, we present a novel end-to-end architecture Mask-Ranking Network(MRNet) for semi-supervised video object segmentation, which take both the advantage of propagation-based methods and matching-based methods and avoiding online fine-tuning. A Mask-Ranking Module is proposed to make a internal guidance for the current frame processing. During the inference, the most conductive information will be selected from the past frames to refine the current

Fig. 10: The qualitative results of our MRNet on DAVIS 2016 and DAVIS 2017 validation set.

segmentation. Experiments on DAVIS 2016 and DAVIS 2017 dataset demonstrate that our MRNet achieves state-of-the-art performance. Overall, MRNet is a practical useful method for video object segmentation.

The future research direction of this problem can be explored from multiple aspects. First, in our MRNet, we calculate the pixel-wise similarity using cosine similarity. There may be other effective similarity measurement methods between frames. Moreover, we measure the similarity for specific candidate frames(every 5 frames). The selection of candidate frames is also a problem worthy of further study. Finally, the mechanism of dynamically exploring long-term similarities between frames can be applied to other video processing tasks. We hope our method can serve as a solid baseline for future research.

Acknowledgements. This work was supported by the National Key Research and Development Program of China (2018YFE0203900), National Science Foundation of China (U1733111, U19A2052), and Sichuan Science and Technology Achievement Transformation project (2020ZHCG0015).

References

1. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: Computer Vision - ECCV 2010 - 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part V. (2010)
2. Grundmann, M., Kwatra, V., Mei, H., Essa, I.: Efficient hierarchical graph-based video segmentation. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. (2010)
3. Tokmakov, P., Alahari, K., Schmid, C.: Learning motion patterns in videos. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 531–539
4. Wang, W., Shen∗, J., Porikli, F.: Saliency-aware geodesic video object segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
5. Anna, Khoreva, Rodrigo, Benenson, Eddy, Ilg, Thomas, Brox, Bernt, Schiele: Lucid data dreaming for video object segmentation. International Journal of Computer Vision (2019)
6. Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixe, L., Cremers, D., Van Gool, L.: One-shot video object segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
7. Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H.: Segflow: Joint learning for video object segmentation and optical flow. In: The IEEE International Conference on Computer Vision (ICCV). (2017)
8. Ci, H., Wang, C., Wang, Y.: Video object segmentation by learning location-sensitive embeddings. (2018)
9. Hu, P., Wang, G., Kong, X., Kuen, J., Tan, Y.: Motion-guided cascaded refinement network for video object segmentation. IEEE Transactions on Pattern Analysis & Machine Intelligence (2019) 1–1
10. Khoreva, A., Perazzi, F., Benenson, R., Schiele, B., Sorkine-Hornung, A.: Learning video object segmentation from static images. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
11. Maninis, K.K., Caelles, S., Chen, Y., Pont-Tuset, J., Leal-Taixe, L., Cremers, D., Van Gool, L.: Video object segmentation without temporal information. IEEE Transactions on Pattern Analysis & Machine Intelligence (2018) 1–1
12. Voigtlaender, P., Leibe, B.: Online adaptation of convolutional neural networks for video object segmentation. arXiv preprint arXiv:1706.09364 (2017)
13. Jang, W.D., Kim, C.S.: Online video object segmentation via convolutional trident network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
14. Johnander, J., Danelljan, M., Brissman, E., Khan, F.S., Felsberg, M.: A generative appearance model for end-to-end video object segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2019)
15. Cheng, J., Tsai, Y.H., Hung, W.C., Wang, S., Yang, M.H.: Fast and accurate online video object segmentation via tracking parts. (2018)
16. Ventura, C., Bellver, M., Girbau, A., Salvador, A., Marques, F., Giro-i Nieto, X.: Rvos: End-to-end recurrent network for video object segmentation. (2019)
17. Robinson, A., Lawin, F.J., Danelljan, M., Khan, F.S., Felsberg, M.: Learning fast and robust target models for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 9384–9393

18. Chen, X., Li, Z., Yuan, Y., Yu, G., Shen, J., Qi, D.: State-aware tracker for real-time video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 9384–9393
19. Sun, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: in Advances in Neural Information Processing Systems. (2014)
20. Liu, Y., Jiang, P.T., Petrosyan, V., Li, S.J., Bian, J., Zhang, L., Cheng, M.M.: Del: Deep embedding learning for efficient image segmentation. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, International Joint Conferences on Artificial Intelligence Organization (2018) 864–870
21. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. (2017)
22. Chen, B., Deng, W.: Deep embedding learning with adaptive large margin n-pair loss for image retrieval and clustering. Pattern Recognition **93** (2019) 353–364
23. Guo, H., Wang, J., Gao, Y., Li, J., Lu, H.: Multi-view 3d object retrieval with deep embedding network. IEEE Transactions on Image Processing **25** (2016) 5526–5537
24. Li, Z., Tang, J., Mei, T.: Deep collaborative embedding for social image understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence **41** (2019) 2070–2083
25. Chen, Y., Pont-Tuset, J., Montes, A., Van Gool, L.: Blazingly fast video object segmentation with pixel-wise metric learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 1189–1198
26. Hu, Y.T., Huang, J.B., Schwing, A.G.: Videomatch: Matching based video object segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 54–70
27. Wang, Z., Xu, J., Liu, L., Zhu, F., Shao, L.: Ranet: Ranking attention network for fast video object segmentation. In: Proceedings of the IEEE international conference on computer vision. (2019) 3978–3987
28. Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Leibe, B., Chen, L.C.: Feelvos: Fast end-to-end embedding learning for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 9481–9490
29. Zhang, Y., Wu, Z., Peng, H., Lin, S.: A transductive approach for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 6949–6958
30. Johnander, J., Danelljan, M., Brissman, E., Khan, F.S., Felsberg, M.: A generative appearance model for end-to-end video object segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 8945–8954
31. Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S., Ling, H.: Learning unsupervised video object segmentation through visual attention. (2019)
32. Bao, L., Wu, B., Liu, W.: Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
33. Hu, Y., Huang, J., Schwing, A.G.: Maskrnn: Instance level video object segmentation. (2017) 325–334
34. Wug Oh, S., Lee, J.Y., Sunkavalli, K., Joo Kim, S.: Fast video object segmentation by reference-guided mask propagation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 7376–7385

35. Luiten, J., Voigtlaender, P., Leibe, B.: Premvos: Proposal-generation, refinement and merging for video object segmentation (2018)
36. Ci, H., Wang, C., Wang, Y.: Video object segmentation by learning location-sensitive embeddings. In: The European Conference on Computer Vision (ECCV). (2018)
37. Li, X., Loy, C.C.: Video object segmentation with joint re-identification and attention-aware mask propagation. In Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., eds.: The European Conference on Computer Vision (ECCV). (2018) 93–110
38. Oh, S.W., Lee, J., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). (2019) 9225–9234
39. Xu, S., Liu, D., Bao, L., Liu, W., Zhou, P.: Mhp-vos: Multiple hypotheses propagation for video object segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 314–323
40. Jampani, V., Gadde, R., Gehler, P.V.: Video propagation networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
41. Yang, L., Wang, Y., Xiong, X., Yang, J., Katsaggelos, A.K.: Efficient video object segmentation via network modulation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 6499–6507
42. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 770–778
43. Pinheiro, P.O., Lin, T.Y., Collobert, R., Dollár, P.: Learning to refine object segments. In: European conference on computer vision, Springer (2016) 75–91
44. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European conference on computer vision, Springer (2016) 630–645
45. Perazzi, F., Pont-Tuset, J., Mcwilliams, B., Gool, L.V., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
46. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
47. Werbos, P., J.: Backpropagation through time: what it does and how to do it. Proceedings of the IEEE **78** (1990) 1550–1560
48. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. Computer Science (2014)